

On the Doubt about Margin Explanation of Boosting

Wei Gao, Zhi-Hua Zhou*

*National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China*

Abstract

Margin theory provides one of the most popular explanations to the success of **AdaBoost**, where the central point lies in the recognition that *margin* is the key for characterizing the performance of **AdaBoost**. This theory has been very influential, e.g., it has been used to argue that **AdaBoost** usually does not overfit since it tends to enlarge the margin even after the training error reaches zero. Previously the *minimum margin bound* was established for **AdaBoost**, however, Breiman [10] pointed out that maximizing the minimum margin does not necessarily lead to a better generalization. Later, Reyzin and Schapire [34] emphasized that the margin distribution rather than minimum margin is crucial to the performance of **AdaBoost**. In this paper, we show that previous margin bounds are special cases of the *kth margin bound*, and none of them is really based on the whole margin distribution. Then, we improve the empirical Bernstein bound given by Maurer and Pontil [28]. Based on this result, we defend the margin-based explanation against Breiman's doubt by proving a new generalization error bound that considers exactly the same factors as Schapire et al. [35] but is uniformly tighter than Breiman [10]'s bound. We also provide a lower bound for generalization error of voting classifiers, and by incorporating factors such as average margin and variance, we present a generalization error bound that is heavily related to the whole margin distribution. Finally, we provide empirical evidence to verify our theory.

Key words: classification, AdaBoost, generalization, overfitting, margin

*Corresponding author. Email: zhouzh@nju.edu.cn

1. Introduction

The **AdaBoost** algorithm [18, 19], which aims to construct a “strong” classifier by combining some “weak” learners (slightly better than random guess), has been one of the most influential classification algorithms [14, 38], and it has exhibited excellent performance both on benchmark datasets and real applications [5, 16].

Many studies are devoted to understanding the mysteries behind the success of **AdaBoost**, among which the margin theory proposed by Schapire et al. [35] has been very influential. For example, **AdaBoost** often tends to be empirically resistant (but not completely) to overfitting [9, 17, 32], i.e., the generalization error of the combined learner keeps decreasing as its size becomes very large and even after the training error has reached zero; it seems violating the Occam’s razor [8], i.e., the principle that less complex classifiers should perform better. This remains one of the most famous mysteries of **AdaBoost**. The margin theory provides the most intuitive and popular explanation to this mystery, that is: **AdaBoost** tends to improve the margin even after the error on training sample reaches zero.

However, Breiman [10] raised serious doubt on the margin theory by designing **arc-gv**, a boosting-style algorithm. This algorithm is able to maximize the *minimum margin* over the training data, but its generalization error is high on empirical datasets. Thus, Breiman [10] concluded that the margin theory for **AdaBoost** failed. Breiman’s argument was backed up with a minimum margin bound, which is tighter than the generalization bound given by Schapire et al. [35], and a lot of experiments. Later, Reyzin and Schapire [34] found that there were flaws in the design of experiments: Breiman used CART trees [12] as base learners and fixed the number of leaves for controlling the complexity of base learners. However, Reyzin and Schapire [34] found that the trees produced by **arc-gv** were usually much deeper than those produced by **AdaBoost**. Generally, for two trees with the same number of leaves, the deeper one is with a larger complexity because more judgements are needed for making a prediction. Therefore, Reyzin and Schapire [34] concluded that Breiman’s observation was biased due to the poor control of model complexity. They repeated the experiments by using decision stumps for base learners, considering that decision stump has only one leaf and thus with a fixed complexity, and observed that though **arc-gv** produced a larger minimum margin, its margin distribution was quite poor. Nowadays, it is well-accepted that the margin distribution is crucial to relate margin to the generalization

performance of **AdaBoost**. To support the margin theory, Wang et al. [37] presented a tighter bound in term of $Emargin$, which was believed to be relevant to margin distribution.

In this paper, we show that the minimum margin and $Emargin$ are special cases of the k th margin, and all the previous margin bounds are single margin bounds that are not really based on the whole margin distribution. Then, we present a new empirical Bernstein bound, which slightly improves the bound in [28] but with different proof skills. Based on this result, we prove a new generalization error bound for voting classifier, which considers exactly the same factors as Schapire et al. [35], but is uniformly tighter than the bounds of Schapire et al. [35] and Breiman [10]. Therefore, we defend the margin-based explanation against Breiman’s doubt. Furthermore, we present a lower generalization error bound for voting classifiers, and by incorporating other factors such as average margin and variance, we prove a generalization error bound which is heavily relevant to the whole margin distribution. Finally, we make a comprehensive empirical comparisons between **AdaBoost** and **arc-gv**, and find that **AdaBoost** has better performance than but dose not absolutely outperform **arc-gv**, which verifies our theory completely.

The rest of this paper is organized as follows. We begin with some notations and background in Sections 2 and 3, respectively. Then, we prove the k th margin bound and discuss on its relation to previous bounds in Section 4. Our main results are presented in Section 5, and detailed proofs are provided in Section 6. We give empirical evidence in Section 7 and conclude this paper in Section 8.

2. Notations

Let \mathcal{X} and \mathcal{Y} denote an input space and output space, respectively. For simplicity, we focus on binary classification problems, i.e., $\mathcal{Y} = \{+1, -1\}$. Denote by D an (unknown) underlying probability distribution over the product space $\mathcal{X} \times \mathcal{Y}$. A training sample with size m

$$S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$$

is drawn independently and identically (i.i.d) according to distribution D . We use $\Pr_D[\cdot]$ to refer as the probability with respect to D , and $\Pr_S[\cdot]$ to denote the probability with respect to uniform distribution over the sample S . Similarly, we use $E_D[\cdot]$ and $E_S[\cdot]$ to denote the expected values, respectively. For an integer $m > 0$, we set $[m] = \{1, 2, \dots, m\}$.

The Bernoulli Kullback-Leiler (or KL) divergence is defined as

$$KL(q||p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \text{ for } 0 \leq p, q \leq 1.$$

For a fixed q , we can easily find that $KL(q||p)$ is a monotone increasing function for $q \leq p < 1$, and thus, the inverse of $KL(q||p)$ for the fixed q is given by

$$KL^{-1}(q; u) = \inf_w \{w : w \geq q \text{ and } KL(q||w) \geq u\}.$$

Let \mathcal{H} be a hypothesis space. Throughout this paper, we restrain \mathcal{H} to be finite, and similar consideration can be made to the case when \mathcal{H} has finite VC-dimension. We denote by

$$\mathcal{A} = \left\{ \frac{i}{|\mathcal{H}|} : i \in [|\mathcal{H}|] \right\}.$$

A base learner $h \in \mathcal{H}$ is a function which maps a distribution over $\mathcal{X} \times \mathcal{Y}$ onto a function $h : \mathcal{X} \rightarrow \mathcal{Y}$. Let $\mathcal{C}(\mathcal{H})$ denote the convex hull of \mathcal{H} , i.e., a voting classifier $f \in \mathcal{C}(\mathcal{H})$ is of the following form

$$f = \sum \alpha_i h_i \text{ with } \sum \alpha_i = 1 \text{ and } \alpha_i \geq 0.$$

For $N \geq 1$, denote by $\mathcal{C}_N(\mathcal{H})$ the set of unweighted averages over N elements from \mathcal{H} , that is

$$\mathcal{C}_N(\mathcal{H}) = \left\{ g : g = \sum_{j=1}^N \frac{h_j}{N}, h_j \in \mathcal{H} \right\}. \quad (1)$$

For voting classifier $f \in \mathcal{C}(\mathcal{H})$, we can associate with a distribution over \mathcal{H} by using the coefficients $\{\alpha_i\}$, denoted by $\mathcal{Q}(f)$. For convenience, $g \in \mathcal{C}_N(\mathcal{H}) \sim \mathcal{Q}(f)$ implies $g = \sum_{j=1}^N h_j/N$ where $h_j \sim \mathcal{Q}(f)$.

For an instance (x, y) , the *margin* with respect to the voting classifier $f = \sum \alpha_i h_i(x)$ is defined as $yf(x)$; in other words,

$$yf(x) = \sum_{i: y=h_i(x)} \alpha_i - \sum_{i: y \neq h_i(x)} \alpha_i,$$

which shows the difference between the weights of base learners that classify (x, y) correctly and the weights of base learners that misclassify (x, y) . Therefore, margin can be viewed as a measure of the confidence of the classification. Given a sample $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, we denote by $\hat{y}_1 f(\hat{x}_1)$ the *minimum margin* and $E_S[yf(x)]$ the *average margin*, which are defined respectively as follows:

$$\hat{y}_1 f(\hat{x}_1) = \min_{i \in [m]} \{y_i f(x_i)\} \quad \text{and} \quad E_S[yf(x)] = \sum_{i=1}^m \frac{y_i f(x_i)}{m}.$$

Algorithm 1 A unified description of AdaBoost and arc-gv

Input: Sample $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ and the number of iterations T .

Initialization: $D_1(i) = 1/m$.

for $t = 1$ to T **do**

1. Construct base learner $h_t: \mathcal{X} \rightarrow \mathcal{Y}$ using the distribution D_t .
2. Choose α_t .
3. Update

$$D_{t+1}(i) = D_t(i) \exp(-\alpha_t y_i h_t(x_i)) / Z_t,$$

where Z_t is a normalization factor (such that D_{t+1} is a distribution).

end for

Output: The final classifier $\text{sgn}[f(x)]$, where

$$f(x) = \sum_{t=1}^T \frac{\alpha_t}{\sum_{t=1}^T \alpha_t} h_t(x).$$

3. Background

In statistical community, great efforts have been devoted to understanding how and why AdaBoost works. Friedman et al. [20] made an important stride by viewing AdaBoost as a stagewise optimization and relating it to fitting an additive logistic regression model. Various new boosting-style algorithms were developed by performing a gradient decent optimization of some potential loss functions [13, 26, 33]. Based on this optimization view, some boosting-style algorithms and their variants have been shown to be Bayes's consistent under different settings [3, 4, 7, 11, 22, 25, 31, 39]. However, these theories can not be used to explain the resistance of AdaBoost to overfitting, and some statistical views have been questioned seriously by Mease and Wyner [30] with empirical evidences. In this paper, we focus on the margin theory.

Algorithm 1 provides a unified description of AdaBoost and arc-gv. The only difference between them lies in the choice of α_t . In AdaBoost, α_t is chosen by

$$\alpha_t = \frac{1}{2} \ln \frac{1 + \gamma_t}{1 - \gamma_t},$$

where $\gamma_t = \sum_{i=1}^m D_t(i) y_i h_t(x_i)$ is called the *edge* of h_t , which is an affine transformation of the error rate of $h_t(x)$. However, **Arc-gv** sets α_t in a different way. Denote by ρ_t the minimum margin of the voting classifier of round $t - 1$, that is,

$$\rho_t = \hat{y}_1 f_t(\hat{x}_1) \text{ with } \rho_1 = 0$$

where

$$f_t = \sum_{s=1}^{t-1} \frac{\alpha_s}{\sum_{s=1}^{t-1} \alpha_s} h_s(x).$$

Then, **Arc-gv** sets α_t as to be

$$\alpha_t = \frac{1}{2} \ln \frac{1 + \gamma_t}{1 - \gamma_t} - \frac{1}{2} \ln \frac{1 + \rho_t}{1 - \rho_t}.$$

Schapire et al. [35] first proposed the margin theory for **AdaBoost** and upper bounded the generalization error as follows:

Theorem 1 [35] *For any $\delta > 0$ and $\theta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size m , every voting classifier f satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \leq \Pr_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}} \left(\frac{\ln m \ln |\mathcal{H}|}{\theta^2} + \ln \frac{1}{\delta}\right)^{1/2}\right).$$

Breiman [10] provided the minimum margin bound for **arc-gv** by Theorem 2 with our notations.

Theorem 2 [10] *If*

$$\theta = \hat{y}_1 f(\hat{x}_1) > 4\sqrt{\frac{2}{|\mathcal{H}|}} \text{ and } R = \frac{32 \ln 2 |\mathcal{H}|}{m\theta^2} \leq 2m,$$

then, for any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size m , every voting classifier f satisfies the following bound:

$$\Pr_D[yf(x) < 0] \leq R\left(\ln(2m) + \ln \frac{1}{R} + 1\right) + \frac{1}{m} \ln \frac{|\mathcal{H}|}{\delta}.$$

Empirical results show that **arc-gv** probably generates a larger minimum margin but with higher generalization error, and Breiman's bound is $O(\frac{\ln m}{m})$, tighter than $O(\sqrt{\frac{\ln m}{m}})$ in Theorem 1. Thus,

Breiman cast serious doubt on margin theory. To support the margin theory, [37] presented a tighter bound in term of Wang et al. *Emargin* by Theorem 3, which was believed to be related to margin distribution. Notice that the factors considered by Wang et al. [37] are different from that considered by Schapire et al. [35] and Breiman [10].

Theorem 3 [37] *For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of the sample S with size m , every voting classifier f satisfying the following bound:*

$$\Pr_D[yf(x) < 0] \leq \frac{\ln |\mathcal{H}|}{m} + \inf_{q \in \{0, \frac{1}{m}, \dots, 1\}} KL^{-1}(q; u[\hat{\theta}(q)]),$$

where

$$u[\hat{\theta}(q)] = \frac{1}{m} \left(\frac{8 \ln |\mathcal{H}|}{\hat{\theta}^2(q)} \ln \frac{2m^2}{\ln |\mathcal{H}|} + \ln |\mathcal{H}| + \ln \frac{m}{\delta} \right)$$

and $\hat{\theta}(q) = \sup \{ \theta \in (\sqrt{8/|\mathcal{H}|}, 1] : \Pr_S[yf(x) \leq \theta] \leq q \}$.

Instead of the whole function space, much work developed margin-based data-dependent bounds for generalization error, e.g., empirical cover number [36], empirical fat-shattering dimension [1], Rademacher and Gaussian complexities [23, 24]. Some of these bounds are proven to be sharper than Theorem 1, but it is difficult, or even impossible, to directly show that these bounds are sharper than the minimum bound of Theorem 2, and fail to explain the resistance of **AdaBoost** to overfitting.

4. None Margin Distribution Bound

Given a sample S of size m , we define the k th margin $\hat{y}_k f(\hat{x}_k)$ as the k th smallest margin over sample S , i.e., the k th smallest value in $\{y_i f(x_i), i \in [m]\}$. The following theorem shows that the k th margin can be used to measure the performance of a voting classifier, whose proof is deferred in Section 6.1.

Theorem 4 *For any $\delta > 0$ and $k \in [m]$, if $\theta = \hat{y}_k f(\hat{x}_k) > \sqrt{8/|\mathcal{H}|}$, then with probability at least $1 - \delta$ over the random choice of sample with size m , every voting classifier f satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \leq \frac{\ln |\mathcal{H}|}{m} + KL^{-1}\left(\frac{k-1}{m}; \frac{q}{m}\right) \quad (2)$$

where

$$q = \frac{8 \ln(2|\mathcal{H}|)}{\theta^2} \ln \frac{2m^2}{\ln |\mathcal{H}|} + \ln |\mathcal{H}| + \ln \frac{m}{\delta}.$$

Especially, when k is constant with $m > 4k$, we have

$$\Pr_D[yf(x) < 0] \leq \frac{\ln |\mathcal{H}|}{m} + \frac{2}{m} \left(\frac{8 \ln(2|\mathcal{H}|)}{\theta^2} \ln \frac{2m^2}{\ln |\mathcal{H}|} + \ln |\mathcal{H}| + \ln \frac{km^{k-1}}{\delta} \right). \quad (3)$$

It is interesting to study the relation between Theorem 4 and previous results, especially for Theorems 2 and 3. It is straightforward to get a result similar to Breiman's minimum margin bound in Theorem 2, by setting $k = 1$ in Eqn. (3):

Corollary 1 *For any $\delta > 0$, if $\theta = \hat{y}_1 f(\hat{x}_1) > \sqrt{8/|\mathcal{H}|}$, then with probability at least $1 - \delta$ over the random choice of sample S with size m , every voting classifier f satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \leq \frac{\ln |\mathcal{H}|}{m} + \frac{2}{m} \left(\frac{8 \ln(2|\mathcal{H}|)}{\theta^2} \ln \frac{2m^2}{\ln |\mathcal{H}|} + \ln \frac{|\mathcal{H}|}{\delta} \right).$$

Notice that when k is a constant, the bound in Eqn. (3) is $O(\ln m/m)$ and the only difference lies in the coefficient. Thus, there is no essential difference to select constant k th margin (such as the 2nd margin, the 3rd margin, etc.) to measure the confidence of classification for large-size sample.

Based on Theorem 4, it is also not difficult to get a result similar to the Emargin bound in Theorem 3 as follows:

Corollary 2 *For any $\delta > 0$, if $\theta_k = \hat{y}_k f(\hat{x}_k) > \sqrt{8/|\mathcal{H}|}$, then with probability at least $1 - \delta$ over the random choice of the sample S with size m , every voting classifier f satisfying the following bound:*

$$\Pr_D[yf(x) < 0] \leq \frac{\ln |\mathcal{H}|}{m} + \inf_{k \in [m]} KL^{-1} \left(\frac{k-1}{m}; \frac{q}{m} \right)$$

where

$$q = \frac{8 \ln(2|\mathcal{H}|)}{\theta_k^2} \ln \frac{2m^2}{\ln |\mathcal{H}|} + \ln |\mathcal{H}| + \ln \frac{m}{\delta}.$$

From Corollary 2, we can easily understand that the Emargin bound ought to be tighter than the minimum margin bound because the former takes the infimum range over $k \in [m]$ while the latter focuses only on the minimum margin.

In summary, the preceding analysis reveals that both the minimum margin and Emargin are special cases of the k th margin; neither of them succeeds in relating margin distribution to the generalization performance of **AdaBoost**.

5. Main Results

We begin with the following empirical Bernstein bound, which is crucial for our main theorems:

Theorem 5 *For any $\delta > 0$, and for i.i.d random variables Z, Z_1, Z_2, \dots, Z_m with $Z \in [0, 1]$ and $m \geq 4$, the followings hold with probability at least $1 - \delta$*

$$E[Z] - \frac{1}{m} \sum_{i=1}^m Z_i \leq \sqrt{\frac{2\hat{V}_m \ln(2/\delta)}{m}} + \frac{7 \ln(2/\delta)}{3m}, \quad (4)$$

$$E[Z] - \frac{1}{m} \sum_{i=1}^m Z_i \geq -\sqrt{\frac{2\hat{V}_m \ln(2/\delta)}{m}} - \frac{7 \ln(2/\delta)}{3m}, \quad (5)$$

where $\hat{V}_m = \sum_{i < j} (Z_i - Z_j)^2 / 2m(m-1)$.

It is noteworthy that the bound in Eqn. (4) is similar to but improves slightly the bound of Maurer and Pontil [28, Theorem 4], and we also present a lower bound as shown in Eqn. (5). This proof is deferred to Section 6.2, which is simple, straightforward and different from [28].

We now present our first main theorem:

Theorem 6 *For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size $m \geq 4$, every voting classifier f satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \leq \frac{2}{m} + \inf_{\theta \in (0, 1]} \left[\Pr_S[yf(x) < \theta] + \frac{7\mu + 3\sqrt{2\mu}}{3m} + \sqrt{\frac{2\mu}{m} \Pr_S[yf(x) < \theta]} \right],$$

where

$$\mu = \frac{8}{\theta^2} \ln m \ln(2|\mathcal{H}|) + \ln \frac{2|\mathcal{H}|}{\delta}.$$

This proof is based on the techniques developed by Schapire et al. [35], and the main difference is that we utilize the empirical Bernstein bound of Eqn. (4) in Theorem 5 for the derivation of generalization error. The detailed proof is deferred to Section 6.3.

It is noteworthy that Theorem 6 shows that the generalization error can be bounded in term of the empirical margin distribution $\Pr_S[yf(x) \leq \theta]$, the training sample size and the hypothesis complexity; in other words, this bound considers exactly the same factors as Schapire et al. [35] in Theorem 1. However, the following corollary shows that, the bound in Theorem 6 is tighter than the bound of Schapire et al. [35] in Theorem 1, as well as the minimum margin bound of Breiman [10] in Theorem 2.

Corollary 3 *For any $\delta > 0$, if the minimum margin $\theta_1 = \hat{y}_1 f(\hat{x}_1) > 0$ and $m \geq 4$, then we have*

$$\inf_{\theta \in (0,1]} \left[\Pr_S[yf(x) < \theta] + \frac{7\mu + 3\sqrt{2\mu}}{3m} + \sqrt{\frac{2\mu}{m} \Pr_S[yf(x) < \theta]} \right] \leq \frac{7\mu_1 + 3\sqrt{2\mu_1}}{3m}, \quad (6)$$

where $\mu = 8 \ln m \ln(2|\mathcal{H}|)/\theta^2 + \ln(2|\mathcal{H}|/\delta)$ and $\mu_1 = 8 \ln m \ln(2|\mathcal{H}|)/\theta_1^2 + \ln(2|\mathcal{H}|/\delta)$; moreover, if the followings hold

$$\theta_1 = \hat{y}_1 f(\hat{x}_1) > 4\sqrt{\frac{2}{|\mathcal{H}|}} \quad (7)$$

$$R = \frac{32 \ln 2|\mathcal{H}|}{m\theta_1^2} \leq 2m \quad (8)$$

$$m \geq \max \left\{ 4, \exp \left(\frac{\theta_1^2}{4 \ln(2|\mathcal{H}|)} \ln \frac{|\mathcal{H}|}{\delta} \right) \right\}, \quad (9)$$

then we have

$$\begin{aligned} \frac{2}{m} + \inf_{\theta \in (0,1]} \left[\Pr_S[yf(x) < \theta] + \frac{7\mu + 3\sqrt{2\mu}}{3m} + \sqrt{\frac{2\mu}{m} \Pr_S[yf(x) < \theta]} \right] \\ \leq R \left(\ln(2m) + \ln \frac{1}{R} + 1 \right) + \frac{1}{m} \ln \frac{|\mathcal{H}|}{\delta}. \end{aligned} \quad (10)$$

This proof is deferred to Section 6.4. From Eqn. (6), we can see clearly that the bound of Theorem 6 is $O(\ln m/m)$, uniformly tighter than the bound of Schapire et al. [35] in Theorem 1. In fact, we could also guarantee that bound of Theorem 6 is $O(\ln m/m)$ even under weaker condition that $\hat{y}_k f(\hat{x}_k) > 0$ for some $k \leq O(\ln m)$. It is also noteworthy Eqns. (7) and (8) are used here to guarantee the conditions of Theorem 2, and Eqn. (10) shows that the bound of Theorem 6 is tighter than Breiman's minimum margin bound of Theorem 2 for large-size sample.

Breiman [10] doubted the margin theory because of two recognitions: i) the minimum margin bound of Breiman [10] is tighter than the margin distribution bound of Schapire et al. [35], and therefore, the minimum margin is more essential than margin distribution to characterize the generalization performance; ii) **arc-gv** maximizes the minimum margin, but demonstrates worse performance than **AdaBoost** empirically. However, our result shows that the margin distribution bound in Theorem 1 can be greatly improved so that it is tighter than the minimum margin bound, and therefore, it is natural that **AdaBoost** outperforms **arc-gv** empirically on some datasets; in a word, our results provide a complete answer to Breiman's doubt on margin theory.

We can also give a lower bound for generalization error as follows:

Theorem 7 *For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size $m \geq 4$, every voting classifier f satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \geq \sup_{\theta \in (0,1]} \left[\Pr_S[yf(x) < -\theta] - \sqrt{\frac{2\mu}{m} \Pr_S[yg(x) < 0]} - \frac{7\mu + 3\sqrt{2\mu}}{3m} \right] - \frac{2}{m}$$

where $\mu = 8 \ln m \ln(2|\mathcal{H}|)/\theta^2 + \ln(2|\mathcal{H}|/\delta)$.

The proof is based on Eqn. (5) in Theorem 5 and we defer it to Section 6.5. We now introduce the second main result as follows:

Theorem 8 *For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size $m \geq 4$, every voting classifier f satisfies the following bound:*

$$\Pr_D[yf(x) < 0] \leq \frac{1}{m^{50}} + \inf_{\theta \in (0,1]} \left[\Pr_S[yf(x) < \theta] + \frac{\sqrt{6\mu}}{m^{3/2}} + \frac{7\mu}{3m} + \sqrt{\frac{2\mu}{m} \hat{\mathcal{I}}(\theta)} + \exp \left(\frac{-2 \ln m}{(1 - E_S^2[yf(x)] + \theta/9)} \right) \right]$$

where $\mu = 144 \ln m \ln(2|\mathcal{H}|)/\theta^2 + \ln(2|\mathcal{H}|/\delta)$ and $\hat{\mathcal{I}}(\theta) = \Pr_S[yf(x) < \theta] \Pr_S[yf(x) \geq 2\theta/3]$.

It is easy to find in almost all boosting experiments that the average margin $E_S[yf(x)]$ is positive. Thus, the bound of Theorem 8 can be tighter when we enlarge the average margin. The statistics $\hat{\mathcal{I}}(\cdot)$ reflects the margin variance in some sense, and the term including $\hat{\mathcal{I}}(\cdot)$ could be small or even vanished except for a small interval when the variance is small. Similarly to the proof of Eqn. (6), we can show that the bound of Theorem 8 is still $O(\ln m/m)$.

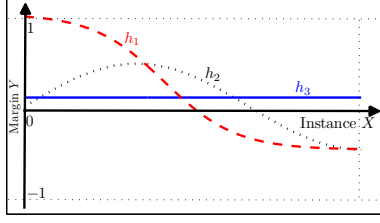


Figure 1: Each curve represents a voting classifier. The X -axis and Y -axis denote instance and margin, respectively, and uniform distribution is assumed on the instance space. The voting classifiers h_1 , h_2 and h_3 have the same average margin but with different generalization error rates: $1/2$, $1/3$ and 0 .

Theorem 8 provides a theoretical support to the suggestion of Reyzin and Schapire [34], that is, the average margin can be used to measure the performance. It is noteworthy that, however, merely considering the average margin is insufficient to bound the generalization error tightly, as shown by the simple example in Figure 1. Indeed, “average” and “variance” are two important statistics for capturing a distribution, and thus, it is reasonable that both the average margin and margin variance are considered in Theorem 8.

6. Proofs

In this section, we provide the detailed proofs for the main theorems and corollaries, and we begin with a series of useful lemmas as follows:

Lemma 1 (Chernoff bound [15]) *Let X, X_1, X_2, \dots, X_m be i.i.d random variables with $X \in [0, 1]$. Then, the followings hold for any $\epsilon > 0$,*

$$\Pr \left[\frac{1}{m} \sum_{i=1}^m X_i \geq E[X] + \epsilon \right] \leq \exp \left(-\frac{m\epsilon^2}{2} \right),$$

$$\Pr \left[\frac{1}{m} \sum_{i=1}^m X_i \leq E[X] - \epsilon \right] \leq \exp \left(-\frac{m\epsilon^2}{2} \right).$$

Lemma 2 (Relative entropy Chernoff bound [21]) *The following holds for $0 < \epsilon < 1$,*

$$\sum_{i=0}^{k-1} \binom{m}{i} \epsilon_N^i (1 - \epsilon_N)^{m-i} \leq \exp \left(-mKL \left(\frac{k-1}{m} \parallel \epsilon \right) \right).$$

Lemma 3 (Bernstein inequalities [6]) *Let X, X_1, X_2, \dots, X_m be i.i.d random variables with $X_i \in [0, 1]$. Then, for any $\delta > 0$, the followings hold with probability at least $1 - \delta$,*

$$E[X] - \frac{1}{m} \sum_{i=1}^m X_i \leq \sqrt{\frac{2V(X) \ln 1/\delta}{m}} + \frac{\ln 1/\delta}{3m}, \quad (11)$$

$$E[X] - \frac{1}{m} \sum_{i=1}^m X_i \geq -\sqrt{\frac{2V(X) \ln 1/\delta}{m}} - \frac{\ln 1/\delta}{3m}, \quad (12)$$

where $V(X)$ denotes the variance $E[(X - E[X])^2]$.

6.1. Proof of Theorem 4

We begin with a lemma as follows:

Lemma 4 *For $f \in \mathcal{C}(\mathcal{H})$ and $g \in \mathcal{C}_N(\mathcal{H})$ chosen i.i.d according to distribution $\mathcal{Q}(f)$. If $\hat{y}_k f(\hat{x}_k) \geq \theta$ and $\hat{y}_k g(\hat{x}_k) \leq \alpha$ with $\theta > \alpha$, then there is an instance (x_i, y_i) in S such that $y_i f(x_i) \geq \theta$ and $y_i g(x_i) \leq \alpha$.*

Proof: There exists a bijection between $\{y_j f(x_j) : j \in [m]\}$ and $\{y_j g(x_j) : j \in [m]\}$ according to the original position in S . Suppose $\hat{y}_k f(\hat{x}_k)$ corresponds to $\hat{y}_l g(\hat{x}_l)$ for some l . If $l \leq k$ then the example (\hat{x}_k, \hat{y}_k) of $\hat{y}_k f(\hat{x}_k)$ is desired; otherwise, except for (\hat{x}_k, \hat{y}_k) of $\hat{y}_k f(\hat{x}_k)$ in S , there are at least $m - k$ elements larger than or equal to θ in $\{y_j f(x_j) : j \in [m] \setminus \{k\}\}$ but at most $m - k - 1$ elements larger than α in $\{y_j g(x_j) : j \in [m] \setminus \{l\}\}$. This completes the proof from the bijection. \square

Proof of Theorem 4: For every $f \in \mathcal{C}(\mathcal{H})$, we can construct a $g \in \mathcal{C}_N(\mathcal{H})$ by choosing N elements i.i.d according to distribution $\mathcal{Q}(f)$, and thus $E_{g \sim \mathcal{Q}(f)}[g] = f$. For $\alpha > 0$, the Chernoff's bound in Lemma 1 gives

$$\begin{aligned} \Pr_D[yf(x) < 0] &= \Pr_{D, \mathcal{Q}(f)}[yf(x) < 0, yg(x) \geq \alpha] + \Pr_{D, \mathcal{Q}(f)}[yf(x) < 0, yg(x) < \alpha] \\ &\leq \exp(-N\alpha^2/2) + \Pr_{D, \mathcal{Q}(f)}[yg(x) < \alpha]. \end{aligned} \quad (13)$$

For any $\epsilon_N > 0$, we consider the following probability:

$$\begin{aligned} &\Pr_{S \sim D^m} \left[\Pr_D[yg(x) < \alpha] > I[\hat{y}_k g(\hat{x}_k) \leq \alpha] + \epsilon_N \right] \\ &\leq \Pr_{S \sim D^m} \left[\hat{y}_k g(\hat{x}_k) > \alpha \mid \Pr_D[yg(x) < \alpha] > \epsilon_N \right] \\ &\leq \sum_{i=0}^{k-1} \binom{m}{i} \epsilon_N^i (1 - \epsilon_N)^{m-i} \end{aligned} \quad (14)$$

where $\hat{y}_k g(\hat{x}_k)$ denotes the k th margin with respect to g . For any k , Eqn. (14) can be bounded by $\exp\left(-mKL\left(\frac{k-1}{m} \parallel \epsilon_N\right)\right)$ from Lemma 2; for constant k with $m > 4k$, we have

$$\sum_{i=0}^{k-1} \binom{m}{i} \epsilon_N^i (1 - \epsilon_N)^{m-i} \leq k(1 - \epsilon_N)^{m/2} \binom{m}{k-1} \leq km^{k-1} (1 - \epsilon_N)^{m/2}.$$

By using the union bound and $|\mathcal{C}_N(\mathcal{H})| \leq |\mathcal{H}|^N$, we have, for any $k \in [m]$,

$$\begin{aligned} & \Pr_{S \sim D^m, g \sim \mathcal{Q}(f)} \left[\exists g \in \mathcal{C}_N(\mathcal{H}), \exists \alpha \in \mathcal{A}, \Pr_D[yg(x) < \alpha] > I[\hat{y}_k g(\hat{x}_k) \leq \alpha] + \epsilon_N \right] \\ & \leq |\mathcal{H}|^{N+1} \exp\left(-mKL\left(\frac{k-1}{m} \parallel \epsilon_N\right)\right). \end{aligned}$$

Setting $\delta_N = |\mathcal{H}|^{N+1} \exp\left(-mKL\left(\frac{k-1}{m} \parallel \epsilon_N\right)\right)$ gives $\epsilon_N = KL^{-1}\left(\frac{k-1}{m}; \frac{1}{m} \ln \frac{|\mathcal{H}|^{N+1}}{\delta_N}\right)$. Thus, with probability at least $1 - \delta_N$ over sample S , for all $f \in \mathcal{C}(\mathcal{H})$ and all $\alpha \in \mathcal{A}$, we have

$$\Pr_D[yg(x) < \alpha] \leq I[\hat{y}_k g(\hat{x}_k) \leq \alpha] + KL^{-1}\left(\frac{k-1}{m}; \frac{1}{m} \ln \frac{|\mathcal{H}|^{N+1}}{\delta_N}\right). \quad (15)$$

Similarly, for constant k , with probability at least $1 - \delta_N$ over sample S , it holds that

$$\Pr_D[yg(x) < \alpha] \leq I[\hat{y}_k g(\hat{x}_k) \leq \alpha] + \frac{2}{m} \ln \frac{km^{k-1} |\mathcal{H}|^{N+1}}{\delta_N}. \quad (16)$$

From $E_{g \sim \mathcal{Q}(f)}[I[\hat{y}_k g(\hat{x}_k) \leq \alpha]] = \Pr_{g \sim \mathcal{Q}(f)}[\hat{y}_k g(\hat{x}_k) \leq \alpha]$, we have, for any $\theta > \alpha$,

$$\Pr_{g \sim \mathcal{Q}(f)}[\hat{y}_k g(\hat{x}_k) \leq \alpha] \leq I[\hat{y}_k f(\hat{x}_k) < \theta] + \Pr_{g \sim \mathcal{Q}(f)}[\hat{y}_k f(\hat{x}_k) \geq \theta, \hat{y}_k g(\hat{x}_k) \leq \alpha]. \quad (17)$$

Notice that the instance (\hat{x}_k, \hat{y}_k) in $\{\hat{y}_i f(\hat{x}_i)\}$ may be different from instance (\hat{x}_k, \hat{y}_k) in $\{\hat{y}_i g(\hat{x}_i)\}$, but from Lemma 4, the last term on the right-hand side of Eqn. (17) can be further bounded by

$$\Pr_{g \sim \mathcal{Q}(f)}[\exists (x_i, y_i) \in S: y_i f(x_i) \geq \theta, y_i g(x_i) \leq \alpha] \leq m \exp(-N(\theta - \alpha)^2/2). \quad (18)$$

Combining Eqns. (13), (15), (17) and (18), we have that with probability at least $1 - \delta_N$ over the sample S , for all $f \in \mathcal{C}(\mathcal{H})$, all $\theta > \alpha$, all $k \in [m]$ but fixed N :

$$\begin{aligned} \Pr_D[yf(x) < 0] & \leq I[\hat{y}_k f(\hat{x}_k) \leq \theta] + m \exp(-N(\theta - \alpha)^2/2) + \exp(-N\alpha^2/2) \\ & \quad + KL^{-1}\left(\frac{k-1}{m}; \frac{1}{m} \ln \frac{|\mathcal{H}|^{N+1} m}{\delta_N}\right). \end{aligned} \quad (19)$$

To obtain the probability of failure for any N at most δ , we select $\delta_N = \delta/2^N$. Setting $\alpha = \frac{\theta}{2} - \frac{\eta}{|\mathcal{H}|} \in \mathcal{A}$ and $N = \frac{8}{\theta^2} \ln \frac{2m^2}{\ln |\mathcal{H}|}$ with $0 \leq \eta < 1$, we have

$$\exp(-N\alpha^2/2) + m \exp(-N(\theta - \alpha)^2/2) \leq 2m \exp(-N\theta^2/8) \leq \ln |\mathcal{H}|/m$$

from the fact $2m > \exp(N/(2|\mathcal{H}|))$ for $\theta > \sqrt{8/|\mathcal{H}|}$. Finally we obtain

$$\Pr[yf(x) < 0] \leq I[\hat{y}_k f(\hat{x}_k) < \theta] + \frac{\ln |\mathcal{H}|}{m} + KL^{-1} \left(\frac{k-1}{m} \parallel \frac{q}{m} \right)$$

where $q = \frac{8 \ln(2|\mathcal{H}|)}{\theta^2} \ln \frac{2m^2}{\ln |\mathcal{H}|} + \ln |\mathcal{H}| + \ln \frac{m}{\delta}$. This completes the proof of Eqn. (2). In a similar manner, we have

$$\Pr[yf(x) < 0] \leq I[\hat{y}_k f(\hat{x}_k) < \theta] + \frac{\ln |\mathcal{H}|}{m} + \frac{2}{m} \left(\frac{8 \ln(2|\mathcal{H}|)}{\theta^2} \ln \frac{2m^2}{\ln |\mathcal{H}|} + \ln |\mathcal{H}| + \ln \frac{km^{k-1}}{\delta} \right),$$

for constant k with $m > 4k$. This completes the proof of Eqn. (3) as desired. \square

6.2. Proof of Theorem 5

For notational simplicity, we denote by $\bar{X} = (X_1, X_2, \dots, X_m)$ a vector of m i.i.d. random variables, and further set $\bar{X}^{k,Y} = (X_1, \dots, X_{k-1}, Y, X_{k+1}, \dots, X_m)$, i.e., the vector with the k th variable X_k in \bar{X} replaced by variable Y . We first introduce some lemmas as follows:

Lemma 5 (McDiarmid Formula [29]) *Let $\bar{X} = (X_1, X_2, \dots, X_m)$ be a vector of m i.i.d. random variables taking values in a set \mathcal{A} . For any $k \in [m]$ and $Y \in \mathcal{A}$, if $|F(\bar{X}) - F(\bar{X}^{k,Y})| \leq c_k$ for $F: \mathcal{A}^m \rightarrow \mathbb{R}$, then the following holds for any $t > 0$*

$$\Pr[F(\bar{X}) - E[F(\bar{X})] \geq t] \leq \exp \left(\frac{-2t^2}{\sum_{k=1}^m c_k^2} \right).$$

Lemma 6 (Theorem 13 [27]) *Let $\bar{X} = (X_1, X_2, \dots, X_m)$ be a vector of m i.i.d. random variables taking values in a set \mathcal{A} . If $F: \mathcal{A}^m \rightarrow \mathbb{R}$ satisfies that*

$$F(\bar{X}) - \inf_{Y \in \mathcal{A}} F(\bar{X}^{k,Y}) \leq 1 \text{ and } \sum_{k=1}^m \left(F(\bar{X}) - \inf_{Y \in \mathcal{A}} F(\bar{X}^{k,Y}) \right)^2 \leq F(\bar{X}),$$

then the following holds for any $t > 0$,

$$\Pr[E[F(\bar{X})] - F(\bar{X}) > t] \leq \exp(-t^2/2E[F(\bar{X})]).$$

Lemma 7 *For two i.i.d random variables X and Y , we have*

$$E[(X - Y)^2] = 2E[(X - E[X])^2] = 2V(X).$$

Proof: This lemma follows from the obvious fact $E[(X - Y)^2] = E(X^2 + Y^2 - 2XY) = 2E[X^2] - 2E^2[X] = 2E[(X - E[X])^2]$. \square

Theorem 9 Let $\bar{X} = (X_1, X_2, \dots, X_m)$ be a vector of $m \geq 4$ i.i.d. random variables with values in $[0, 1]$, and we denote by

$$\hat{V}_m(\bar{X}) = \frac{1}{2m(m-1)} \sum_{i \neq j} (X_i - X_j)^2.$$

Then for any $\delta > 0$, we have

$$\Pr \left[\sqrt{E[\hat{V}_m(\bar{X})]} < \sqrt{\hat{V}_m(\bar{X})} - \sqrt{\frac{\ln 1/\delta}{16m}} \right] \leq \delta, \quad (20)$$

$$\Pr \left[\sqrt{E[\hat{V}_m(\bar{X})]} > \sqrt{\hat{V}_m(\bar{X})} + \sqrt{\frac{2 \ln 1/\delta}{m}} \right] \leq \delta. \quad (21)$$

The bounds in this theorem are tighter than the bounds of [28, Theorem 10], in particularly for Eqn. (20). However, our proof is simple, direct and different from work of Maurer and Pontil.

Proof of Theorem 9 We will utilize Lemmas 5 and 6 to prove Eqns. (20) and (21), respectively.

For Eqn. (20), we first observe that, for any $k \in [m]$,

$$\left| \sqrt{\hat{V}_m(\bar{X})} - \sqrt{\hat{V}_m(\bar{X}^{k,Y})} \right| = \left| \frac{\hat{V}_m(\bar{X}) - \hat{V}_m(\bar{X}^{k,Y})}{\sqrt{\hat{V}_m(\bar{X})} + \sqrt{\hat{V}_m(\bar{X}^{k,Y})}} \right| \leq \frac{1}{2\sqrt{2m}},$$

where we use $\hat{V}_m(\bar{X}), \hat{V}_m(\bar{X}^{k,Y}) \leq 1/2$ from $X_i \in [0, 1]$. By using the Jensen's inequality, we have $E[\sqrt{\hat{V}_m(\bar{X})}] \leq \sqrt{E[\hat{V}_m(\bar{X})]}$ and thus,

$$\Pr \left[\sqrt{E[\hat{V}_m(\bar{X})]} < \sqrt{\hat{V}_m(\bar{X})} - \epsilon \right] \leq \Pr \left[E \left[\sqrt{\hat{V}_m(\bar{X})} \right] < \sqrt{\hat{V}_m(\bar{X})} - \epsilon \right] \leq \exp(-16m\epsilon^2).$$

where the last inequality holds by applying McDiarmid formula in Lemma 5 to $\sqrt{\hat{V}_m}$. Therefore, we complete the proof of Eqn. (20) by setting $\delta = \exp(-16m\epsilon^2)$.

For Eqn. (21), we set $\xi_m(\bar{X}) = m\hat{V}_m(\bar{X})$. For $X_i \in [0, 1]$ and $\xi_m(\bar{X}^{k,Y})$, it is easy to obtain the optimal solution by simple calculation

$$Y^* = \arg \inf_{Y \in [0,1]} [\xi_m(\bar{X}^{k,Y})] = \sum_{i \neq k} \frac{X_i}{m-1},$$

which yields that

$$\xi_m(\bar{X}) - \inf_{Y \in [0,1]} [\xi_m(\bar{X}^{k,Y})] = \frac{1}{m-1} \sum_{i \neq k} (X_i - X_k)^2 - (Y^* - X_k)^2 = \left(X_k - \sum_{i \neq k} \frac{X_i}{m-1} \right)^2.$$

For $X_i \in [0, 1]$, it is obvious that

$$\xi_m(\bar{X}) - \inf_{Y \in [0,1]} [\xi_m(\bar{X}^{k,Y})] \leq 1,$$

and we further have

$$\begin{aligned} \sum_{k=1}^m (\xi_m(\bar{X}) - \inf_{Y \in [0,1]} [\xi_m(\bar{X}^{k,Y})])^2 &= \sum_{k=1}^m \left(X_k - \sum_{i \neq k} \frac{X_i}{m-1} \right)^4 \\ &= \frac{m^5}{(m-1)^4} \frac{1}{m} \sum_{k=1}^m \left(X_k - \sum_{i=1}^m \frac{X_i}{m} \right)^4 \leq \frac{m^5}{(m-1)^4} \left(\frac{1}{m} \sum_{k=1}^m \left(X_k - \sum_{i=1}^m \frac{X_i}{m} \right)^2 \right)^2 \end{aligned} \quad (22)$$

where we use the Jensen's inequality $E[a^4] \leq E^2[a^2]$. From Lemma 7, we have

$$\frac{1}{m} \sum_{k=1}^m \left(X_k - \sum_{i=1}^m \frac{X_i}{m} \right)^2 \leq \frac{1}{2m^2} \sum_{i,k} (X_i - X_k)^2 = \frac{1}{2m^2} \sum_{i \neq k} (X_i - X_k)^2.$$

Substituting the above inequality into Eqn. (22), we have

$$\begin{aligned} \sum_{k=1}^m (\xi_m(\bar{X}) - \inf_{Y \in [0,1]} [\xi_m(\bar{X}^{k,Y})])^2 &\leq \frac{m^3}{4(m-1)^2} \left(\frac{1}{m(m-1)} \sum_{i \neq k} (X_i - X_k)^2 \right)^2 \\ &\leq \frac{m^3}{4(m-1)^2} \frac{1}{m(m-1)} \sum_{i \neq k} (X_i - X_k)^2 \\ &= \frac{m^2}{2(m-1)^2} \xi_m(\bar{X}) \leq \xi_m(\bar{X}) \end{aligned}$$

where the second inequality holds from $\sum_{i \neq k} (X_i - X_k)^2 / m(m-1) \leq 1$ for $X_i \in [0, 1]$ and the last inequality holds from $m \geq 4$. Therefore, for any $t > 0$, the following holds by using Lemma 6 to $\xi_m(\bar{X})$,

$$\Pr[E[\hat{V}_m(\bar{X})] - \hat{V}_m(\bar{X}) > t] = \Pr[E[\xi_m(\bar{X})] - \xi_m(\bar{X}) > mt] \leq \exp \left(\frac{-mt^2}{2E[\hat{V}_m(\bar{X})]} \right).$$

Setting $\delta = \exp(-mt^2/2E[\hat{V}_m(\bar{X})])$ gives

$$\Pr \left[E[\hat{V}_m(\bar{X})] - \hat{V}_m(\bar{X}) > \sqrt{2E[\hat{V}_m(\bar{X})] \ln(1/\delta)/m} \right] \leq \delta$$

which completes the proof of Eq. (21) by using the square-root's inequality and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. \square

Proof of Theorem 5 For i.i.d. random variables $\bar{X} = (X_1, X_2, \dots, X_m)$, we set $\hat{V}_m(\bar{X}) = \sum_{i \neq j} (X_i - X_j)^2 / 2m(m-1)$, and observe that

$$E[\hat{V}_m(\bar{X})] = \frac{1}{2m(m-1)} \sum_{i \neq j} E[(X_i - X_j)^2] = \frac{1}{2m(m-1)} \sum_{i \neq j} 2E[X_i^2] - 2E^2[X_i] = V(X_1),$$

where $V(X_1)$ denotes the variance $V(X_1) = E[(X_1 - E[X_1])^2]$. For any $\delta > 0$, the following holds with probability at least $1 - \delta$ from Eqn. (11),

$$E[X] - \frac{1}{m} \sum_{i=1}^m X_i \leq \sqrt{\frac{2V(X) \ln 1/\delta}{m}} + \frac{\ln 1/\delta}{3m} = \sqrt{\frac{2E[\hat{V}_m(\bar{X})] \ln 1/\delta}{m}} + \frac{\ln 1/\delta}{3m},$$

which completes the proof of Eqn. (4) by combining with Eqn. (21) in a union bound and simple calculations. Similar proof could be made for Eqn. (5). \square

6.3. Proof of Theorem 6

Similarly to the proof of Theorem 4, we have

$$\Pr_D[yf(x) < 0] \leq \exp(-N\alpha^2/2) + \Pr_{D, \mathcal{Q}(f)}[yg(x) < \alpha], \quad (23)$$

for any given $\alpha > 0$, $f \in \mathcal{C}(\mathcal{H})$ and $g \in \mathcal{C}_N(\mathcal{H})$ chosen i.i.d according to $\mathcal{Q}(f)$. Recall that $|\mathcal{C}_N(\mathcal{H})| \leq |\mathcal{H}|^N$. Therefore, for any $\delta_N > 0$, combining union bound with Eqn. (4) in Theorem 5 guarantees that the following holds with probability at least $1 - \delta_N$ over sample S , for any $g \in \mathcal{C}_N(\mathcal{H})$ and $\alpha \in \mathcal{A}$,

$$\Pr_D[yg(x) < \alpha] \leq \Pr_S[yg(x) < \alpha] + \sqrt{\frac{2}{m} \hat{V}_m \ln \left(\frac{2}{\delta_N} |\mathcal{H}|^{N+1} \right)} + \frac{7}{3m} \ln \left(\frac{2}{\delta_N} |\mathcal{H}|^{N+1} \right), \quad (24)$$

where

$$\hat{V}_m = \sum_{i < j} \frac{(I[y_i g(f(x_i)) < \alpha] - I[y_j g(f(x_j)) < \alpha])^2}{2m(m-1)}.$$

Furthermore, we have

$$\sum_{i < j} (I[y_i g(f(x_i)) < \alpha] - I[y_j g(f(x_j)) < \alpha])^2 = m^2 \Pr_S[yg(x) < \alpha] \Pr_S[yg(x) \geq \alpha],$$

which yields that

$$\hat{V}_m = \frac{m}{2m-2} \Pr_S[yg(x) < \alpha] \Pr_S[yg(x) \geq \alpha] \leq \Pr_S[yg(x) < \alpha], \quad (25)$$

for $m \geq 4$. By using Lemma 1 again, the following holds for any $\theta_1 > 0$,

$$\Pr_S[yg(x) < \alpha] \leq \exp(-N\theta_1^2/2) + \Pr_S[yf(x) < \alpha + \theta_1]. \quad (26)$$

Setting $\theta_1 = \alpha = \theta/2$ and combining Eqns. (23), (24), (25) and (26), we have

$$\begin{aligned} \Pr_D[yf(x) < 0] &\leq \Pr_S[yf(x) < \theta] + 2\exp(-N\theta^2/8) \\ &\quad + \frac{7\mu}{3m} + \sqrt{\frac{2\mu}{m} \left(\Pr_S[yf(x) < \theta] + \exp\left(-\frac{N\theta^2}{8}\right) \right)}, \end{aligned}$$

where $\mu = \ln(2|\mathcal{H}|^{N+1}/\delta_N)$. By utilizing the fact $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a \geq 0$ and $b \geq 0$, we further have

$$\sqrt{\frac{2\mu}{m} \left(\Pr_S[yf(x) < \theta] + \exp\left(-\frac{N\theta^2}{8}\right) \right)} \leq \sqrt{\frac{2\mu}{m} \Pr_S[yf(x) < \theta]} + \sqrt{\frac{2\mu}{m} \exp\left(-\frac{N\theta^2}{8}\right)}.$$

Finally, we set $\delta_N = \delta/2^N$ so that the probability of failure for any N will be no more than δ . This theorem follows by setting $N = 8 \ln m / \theta^2$. \square

6.4. Proof of Corollary 3

If the minimum margin $\theta_1 = \hat{y}_1 f(\hat{x}_1) > 0$, then we have $\Pr_S[yf(x) < \theta_1] = 0$ and further get

$$\begin{aligned} &\inf_{\theta \in (0,1]} \left[\Pr_S[yf(x) < \theta] + \frac{7\mu + 3\sqrt{2\mu}}{3m} + \sqrt{\frac{2\mu}{m} \Pr_S[yf(x) < \theta]} \right] \\ &\leq \Pr_S[yf(x) < \theta_1] + \frac{7\mu_1 + 3\sqrt{2\mu_1}}{3m} + \sqrt{\frac{2\mu_1}{m} \Pr_S[yf(x) < \theta_1]} \\ &= \frac{7\mu_1 + 3\sqrt{2\mu_1}}{3m}, \end{aligned} \tag{27}$$

where $\mu_1 = 8 \ln m \ln(2|\mathcal{H}|)/\theta_1^2 + \ln(2|\mathcal{H}|/\delta)$. This gives the proof of Eqn. (6). If $m \geq 4$, then we have

$$\mu_1 \geq \frac{8}{\theta_1^2} \ln m \ln(2|\mathcal{H}|) \geq 5 \text{ leading to } \sqrt{2\mu_1} \leq 2\mu_1/3.$$

Therefore, the following holds by combining Eqn. (27) and the above facts,

$$\begin{aligned} &\frac{2}{m} + \inf_{\theta \in (0,1]} \left[\Pr_S[yf(x) < \theta] + \frac{7\mu + 3\sqrt{2\mu}}{3m} + \sqrt{\frac{3\mu}{m} \Pr_S[yf(x) < \theta]} \right] \\ &\leq \frac{2}{m} + \frac{7\mu_1 + 3\sqrt{2\mu_1}}{3m} \leq \frac{2}{m} + \frac{3\mu_1}{m} = \frac{2}{m} + \frac{24 \ln m}{m\theta_1^2} \ln(2|\mathcal{H}|) + \frac{3}{m} \ln \frac{2|\mathcal{H}|}{\delta} \\ &\leq \frac{8}{m} + \frac{24 \ln m}{m\theta_1^2} \ln(2|\mathcal{H}|) + \frac{3}{m} \ln \frac{|\mathcal{H}|}{\delta} \leq R \left(\ln(2m) + \ln \frac{1}{R} + 1 \right) + \frac{1}{m} \ln \frac{|\mathcal{H}|}{\delta} \end{aligned}$$

where the last inequality holds from the conditions of Eqn. (9) and $8/m < R$. This completes the proof of Eqn. (10). \square

6.5. Proof of Theorem 7

Proof: For any given $\alpha > 0$, $f \in \mathcal{C}(\mathcal{H})$ and $g \in \mathcal{C}_N(\mathcal{H})$ chosen i.i.d according to $\mathcal{Q}(f)$, it holds that from Lemma 1,

$$\Pr_D[yf(x) \geq 0] \leq \Pr_{D, \mathcal{Q}(f)}[yg(x) \geq -\alpha] + \exp(-N\alpha^2/2),$$

which yields

$$\Pr_D[yf(x) < 0] \geq \Pr_{D, \mathcal{Q}(f)}[yg(x) < -\alpha] - \exp(-N\alpha^2/2). \quad (28)$$

Recall that $|\mathcal{C}_N(\mathcal{H})| \leq |\mathcal{H}|^N$. Therefore, for any $\delta_N > 0$, combining union bound with Eqn. (5) in Theorem 5 guarantees that the following holds with probability at least $1 - \delta_N$ over sample S , for any $g \in \mathcal{C}_N(\mathcal{H})$ and $\alpha \in \mathcal{A}$,

$$\Pr_D[yg(x) < -\alpha] \geq \Pr_S[yg(x) < -\alpha] - \sqrt{\frac{2}{m} \hat{V}_m \ln\left(\frac{2}{\delta_N} |\mathcal{H}|^{N+1}\right)} - \frac{7}{3m} \ln\left(\frac{2}{\delta_N} |\mathcal{H}|^{N+1}\right), \quad (29)$$

where

$$\hat{V}_m = \sum_{i < j} \frac{(I[y_i g(f(x_i)) < -\alpha] - I[y_j g(f(x_j)) < -\alpha])^2}{2m^2 - 2m} \leq \Pr_S[yg(x) < -\alpha] \text{ for } m \geq 4.$$

By using Lemma 1 again, it holds holds that,

$$\begin{aligned} \Pr_S[yg(x) < -\alpha] &\leq \Pr_S[yg(x) < 0] + \exp(-N\alpha^2/2), \\ \Pr_S[yg(x) < -\alpha] &\geq \Pr_S[yg(x) < -2\alpha] - \exp(-N\alpha^2/2). \end{aligned}$$

Therefore, combining the above inequalities with Eqns. (28) and (29), we have

$$\begin{aligned} \Pr_D[yf(x) < 0] &\geq \Pr_S[yf(x) < -2\alpha] - 2\exp(-N\alpha^2/2) \\ &\quad - \sqrt{\frac{2 \Pr_S[yg(x) < 0] + 2\exp(-N\alpha^2/2)}{m} \ln\left(\frac{2}{\delta_N} |\mathcal{H}|^{N+1}\right)} - \frac{7}{3m} \ln\left(\frac{2}{\delta_N} |\mathcal{H}|^{N+1}\right) \end{aligned}$$

Set $\theta = 2\alpha$ and $\delta_N = \delta/2^N$ so that the probability of failure for any N will be no more than δ . This theorem follows by using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and setting $N = 8 \ln m / \theta^2$. \square

6.6. Proof of Theorem 8

Our proof is based on a new Bernstein-type bound as follows:

Lemma 8 For $f \in \mathcal{C}(\mathcal{H})$ and $g \in \mathcal{C}_N(\mathcal{H})$ chosen i.i.d according to distribution $\mathcal{Q}(f)$, we have

$$\Pr_{S, g \sim \mathcal{Q}(f)} [yg(x) - yf(x) \geq t] \leq \exp \left(\frac{-Nt^2}{2 - 2E_S^2[yf(x)] + 4t/3} \right).$$

Proof: For $\lambda > 0$, we utilize the Markov's inequality to have

$$\begin{aligned} \Pr_{S, g \sim \mathcal{Q}(f)} [yg(x) - yf(x) \geq t] &= \Pr_{S, g \sim \mathcal{Q}(f)} [(yg(x) - yf(x))N\lambda/2 \geq N\lambda t/2] \\ &\leq \exp \left(-\frac{\lambda Nt}{2} \right) E_{S, g \sim \mathcal{Q}(f)} \left[\exp \left(\frac{\lambda}{2} \sum_{j=1}^N yh_j(x) - yf(x) \right) \right] \\ &= \exp(-\lambda Nt/2) \prod_{j=1}^N E_{S, h_j \sim \mathcal{Q}(f)} [\exp(\lambda(yh_j(x) - yf(x))/2)], \end{aligned}$$

where the last inequality holds from the independence of h_j . Notice that $|yh_j(x) - yf(x)| \leq 2$ from $\mathcal{H} \subseteq \{h: \mathcal{X} \rightarrow \{-1, +1\}\}$. By using Taylor's expansion, we further get

$$\begin{aligned} E_{S, h_j \sim \mathcal{Q}(f)} [\exp(\lambda(yh_j(x) - yf(x))/2)] &\leq 1 + E_{S, h_j \sim \mathcal{Q}(f)} [(yh_j(x) - yf(x))^2](e^\lambda - 1 - \lambda)/4 \\ &= 1 + E_S[1 - (yf(x))^2](e^\lambda - 1 - \lambda)/4 \leq \exp \left((1 - E_S^2[yf(x)])(e^\lambda - 1 - \lambda)/4 \right), \end{aligned}$$

where the last inequality holds from Jensen's inequality and $1 + x \leq e^x$. Therefore, it holds that

$$\Pr_{S, g \sim \mathcal{Q}(f)} [yg(x) - yf(x) \geq t] \leq \exp \left(N(e^\lambda - 1 - \lambda)(1 - E_S^2[yf(x)])/4 - \lambda Nt/2 \right).$$

If $0 < \lambda < 3$, then we could use Taylor's expansion again to have

$$e^\lambda - \lambda - 1 = \sum_{i=2}^{\infty} \frac{\lambda^i}{i!} \leq \frac{\lambda^2}{2} \sum_{i=0}^{\infty} \frac{\lambda^m}{3^m} = \frac{\lambda^2}{2(1 - \lambda/3)}.$$

Now by picking $\lambda = t/(1/2 - E_S^2[yf(x)]/2 + t/3)$, we have

$$-\frac{\lambda t}{2} + \frac{\lambda^2(1 - E_S^2[yf(x)])}{8(1 - \lambda/3)} \leq \frac{-t^2}{2 - 2E_S^2[yf(x)] + 4t/3},$$

which completes the proof as desired. \square

Proof of Theorem 8 This proof is rather similar to the proof of Theorem 6, and we just give main steps. For any $\alpha > 0$ and $\delta_N > 0$, the following holds with probability at least $1 - \delta_N$ over sample S_m ($m \geq 4$),

$$\Pr_D[yf(x) < 0] \leq \Pr_S[yg(x) < \alpha] + \exp \left(-\frac{N\alpha^2}{2} \right) + \sqrt{\frac{2\hat{V}_m^* \ln(\frac{2}{\delta_N} |\mathcal{H}|^{N+1})}{m}} + \frac{7}{3m} \ln(\frac{2}{\delta_N} |\mathcal{H}|^{N+1}),$$

where $\hat{V}_m^* = \Pr_S[yg(x) < \alpha] \Pr_S[yg(x) \geq \alpha]$. For any $\theta_1 > 0$, we use Lemma 1 to obtain

$$\hat{V}_m^* = \Pr_S[yg(x) < \alpha] \Pr_S[yg(x) \geq \alpha] \leq 3 \exp(-N\theta_1^2/2) + \Pr_S[yf(x) < \alpha + \theta_1] \Pr_S[yf(x) > \alpha - \theta_1].$$

From Lemma 8, it holds that

$$\Pr_S[yg(x) < \alpha] \leq \Pr_S[yf(x) < \alpha + \theta_1] + \exp\left(\frac{-N\theta_1^2}{2 - 2E_S^2[yf(x)] + 4\theta_1/3}\right).$$

Let $\theta_1 = \theta/6$, $\alpha = 5\theta/6$, and set $\delta_N = \delta/2^N$ so that the probability of failure for any N will be no more than δ . We complete the proof by setting $N = 144 \ln m/\theta^2$ and simple calculation. \square

7. Empirical Verifications

Though this paper mainly focuses on the theoretical explanation to **AdaBoost**, we also present empirical studies to compare **AdaBoost** and **arc-gv** in terms of their performance so as to verify our theory.

We conduct our experiments on 51 benchmark datasets from the UCI repository [2], which show considerable diversity in size, number of classes, and number and types of attributes. The detailed characteristics are summarized in Table 2, and most of them are investigated by previous researchers. For multi-class datasets, we transform them into two-class datasets by regarding the union of a half number of classes as one meta-class, while the other half as another meta-class, and the partition is selected by making the two meta-classes be with similar sizes. To control the complexity of base learners, we take decision stumps in our experiments as the base learners for both **AdaBoost** and **arc-gv**. On each dataset we run 10 trials of 10-fold cross validation, and the detailed results are summarized in Tables 1.

As shown by previous empirical work [10, 34], we can see clearly from Tables 1 that **AdaBoost** has better performance than **arc-gv**, which also verifies our Corollary 3. On the other hand, it is noteworthy that **AdaBoost** does not absolutely outperform **arc-gv** since the performances of two algorithms are comparable on many datasets. This is because that the bound of Theorem 6 and the minimum margin bound of Theorem 2 are both $O(\ln m/m)$ though former has smaller coefficients.

8. Conclusion

The margin theory provides one of the most intuitive and popular theoretical explanations to **AdaBoost**. It is well-accepted that the margin distribution is crucial for characterizing the performance of **AdaBoost**, and it is desirable to theoretically establish generalization bounds based on margin distribution.

In this paper, we show that previous margin bounds, such as the minimum margin bound and Emargin bound, are all single-margin bounds that do not really depend on the whole margin distribution. Then, we improve slightly the empirical Bernstein bound with different skills. As our main results, we prove a new generalization bound which considers exactly the same factors as Schapire et al. [35] but is uniformly tighter than the bounds of Schapire et al. [35] and Breiman [10], and thus provide a complete answer to Breiman’s doubt on the margin theory. By incorporating other factors such as average margin and variance, we prove another upper bound which is heavily related to the whole margin distribution. Our empirical evidence shows that AdaBoost has better performance than but not absolutely outperform arc-gv, which further confirm our theory.

References

- [1] A. Antos, B. Kégl, T. Linder, and G. Lugosi. Data-dependent margin-based generalization bounds for classification. *Journal of Machine Learning Research*, 3:73–98, 2002.
- [2] A. Asuncion and D. J. Newman. UCI repository of machine learning databases, 2007.
- [3] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.
- [4] P. L. Bartlett and M. Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8:2347–2368, 2007.
- [5] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning*, 36:105–39, 1999.
- [6] S. Bernshtein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

- [7] J. P. Bickel, Y. Ritov, and A. Zakai. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7:705–732, 2006.
- [8] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam’s razor. *Information Processing Letter*, 24(6):377–380, 1987.
- [9] L. Breiman. Arcing algorithm. *Annals of Statistics*, 26:801–849, 1998.
- [10] L. Breiman. Prediction games and arcing classifiers. *Neural Computation*, 11(7):1493–1517, 1999.
- [11] L. Breiman. Some infinity theory for predictor ensembles. Technical Report 577, Statistics Department, University of California, Berkeley, CA, 2000.
- [12] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Chapman & Hall/CRC, Wadsworth, 1984.
- [13] P. Buhlmann and B. Yu. Boosting with l_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.
- [14] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceeding of 23rd International Conference on Machine Learning*, pages 161–168, Pittsburgh, Pennsylvania, 2006.
- [15] H. Chernoff. A measure of asymptotic efficiency of tests of a hypothesis based upon the sum of the observations. *Annals of Mathematical Statistics*, 24:493–507, 1952.
- [16] T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, 40:139–157, 2000.
- [17] H. Drucker and C. Cortes. Boosting decision trees. In D. S. Touretzky, M. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 479–485. MIT Press, Cambridge, MA, 1996.
- [18] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceeding of 13rd International Conference on Machine Learning*, pages 148–156, Bari, Italy, 1996.
- [19] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

- [20] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. with discussions. *Annals of Statistics*, 28(2):337–407, 2000.
- [21] W. Hoeffding. Probability inequalities for sum of bounded random variables. *Journal of American Statistical Society*, 58:13–30, 1963.
- [22] W. Jiang. Process consistency for AdaBoost. *Annals of Statistics*, 32:13–29, 2004.
- [23] L. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30:1–50, 2002.
- [24] L. Koltchinskii and D. Panchenko. Complexities of convex combinations and bounding the generalization error in classification. *Annals of Statistics*, 33:1455–1496, 2005.
- [25] G. Lugosi and N. Vayatis. On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 32:30–55, 2004.
- [26] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean. Boosting algorithms as gradient descent. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 512–518. MIT Press, Cambridge, MA, 1999.
- [27] A. Maurer. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 29(2):121–138, 2006.
- [28] A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, Montreal, Canada, 2009.
- [29] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, Cambridge, UK, 1989.
- [30] D. Mease and A. Wyner. Evidence contrary to the statistical view of boosting with discussion. *Journal of Machine Learning Research*, 9:131–201, 2008.
- [31] I. Mukherjee, C. Rudin, and R. Schapire. The rate of convergence of Adaboost. In *Proceedings of the 24th Annual Conference on Learning Theory*, Budapest, Hungary, 2011.
- [32] J. R. Quinlan. Bagging, boosting, and C4.5. In *Proceeding of 13th National Conference on Artificial Intelligence*, pages 725–730, Portland, OR, 1996.

- [33] G. Rätsch, T. Onoda, and K. R. Müller. Soft margins for Adaboost. *Machine Learning*, 42:287–320, 2001.
- [34] L. Reyzin and R. E. Schapire. How boosting the margin can also boost classifier complexity. In *Proceeding of 23rd International Conference on Machine Learning*, pages 753–760, Pittsburgh, PA, 2006.
- [35] R. Schapire, Y. Freund, P. L. Bartlett, and W. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.
- [36] J. Shawe-Taylor and R. C. Williamson. Generalization performance of classifiers in terms of observed covering numbers. In H. U. Simon P. Fischer, editor, *Proceedings of the 14th European Computational Learning Theory Conference*, pages 153–167, Springer, Berlin, 1999.
- [37] L. W. Wang, M. Sugiyama, C. Yang, Z.-H. Zhou, and J. Feng. A refined margin analysis for boosting algorithms via equilibrium margin. *Journal of Machine Learning Research*, 12:1835–1863, 2011.
- [38] X. Wu and V. Kumar. *The Top Ten Algorithms in Data Mining*. Chapman and Hall/CRC, 2009.
- [39] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004.

Table 1: Accuracy (mean \pm std.) comparisons of AdaBoost and arc-gv on 51 benchmark datasets. The better performance (paired t -test at 95% significance level) is bold. The last line shows the win/tie/loss counts of AdaBoost versus arc-gv.

Dataset	Test error		Dataset	Test error	
	AdaBoost	Arc-gv		AdaBoost	Arc-gv
anneal	0.0047 \pm 0.0066	0.0043 \pm 0.0067	abalone	0.2203 \pm 0.0208	0.2186 \pm 0.0224
artificial	0.3351 \pm 0.0197	0.2666\pm0.0200	auto-m	0.1143 \pm 0.0471	0.1085\pm0.0436
auto	0.0991 \pm 0.0670	0.0996 \pm 0.0667	balance	0.0088 \pm 0.0119	0.0093 \pm 0.0120
breast-w	0.0411 \pm 0.0221	0.0413 \pm 0.0242	car	0.0502 \pm 0.0154	0.0509 \pm 0.0168
cmc	0.2787\pm0.0288	0.2872 \pm 0.0311	colic	0.1905 \pm 0.0661	0.1935 \pm 0.0683
credit-a	0.1368\pm0.0410	0.1622 \pm 0.0405	cylinder	0.2076 \pm 0.0509	0.2070 \pm 0.0570
diabetes	0.2409\pm0.0423	0.2551 \pm 0.0440	german	0.2486\pm0.0372	0.2717 \pm 0.0403
glass	0.2045 \pm 0.0794	0.2113 \pm 0.0848	heart-c	0.1960\pm0.0701	0.2161 \pm 0.0754
heart-h	0.1892\pm0.0623	0.2006 \pm 0.0673	hepatitis	0.1715 \pm 0.0821	0.1798 \pm 0.0848
house-v	0.0471 \pm 0.0333	0.0471 \pm 0.0326	hypo	0.0053 \pm 0.0035	0.0054 \pm 0.0034
ion	0.0721 \pm 0.0432	0.0767 \pm 0.0421	iris	0.0000 \pm 0.0000	0.0000 \pm 0.0000
isolet	0.1270 \pm 0.0113	0.1214\pm0.0116	kr-vs-kp	0.0354 \pm 0.0106	0.0326\pm0.0097
letter	0.1851 \pm 0.0076	0.1778\pm0.0077	lymph	0.1670 \pm 0.0971	0.1690 \pm 0.0972
magic04	0.1555\pm0.0078	0.1578 \pm 0.0077	mfeat-f	0.0445\pm0.0136	0.0471 \pm 0.0143
mfeat-m	0.0990\pm0.0190	0.1048 \pm 0.0200	mush	0.0000 \pm 0.0000	0.0000 \pm 0.0000
musk	0.0916 \pm 0.0413	0.0926 \pm 0.0437	nursery	0.0002 \pm 0.0004	0.0002 \pm 0.0004
optdigits	0.1060 \pm 0.0144	0.1048 \pm 0.0129	page-b	0.0331 \pm 0.0068	0.0325 \pm 0.0062
pendigits	0.0796 \pm 0.0083	0.0788 \pm 0.0081	satimage	0.0565 \pm 0.0083	0.0531\pm0.0080
segment	0.0171 \pm 0.0083	0.0159\pm0.0083	shuttle	0.0010 \pm 0.0001	0.0009\pm0.0001
sick	0.0250 \pm 0.0082	0.0246 \pm 0.0079	solar-f	0.0440\pm0.0171	0.0490 \pm 0.0182
sonar	0.1441\pm0.0697	0.1863 \pm 0.0881	soybean	0.0245 \pm 0.0188	0.0242 \pm 0.0174
spamb	0.0570 \pm 0.0107	0.0553\pm0.0105	spect	0.1256 \pm 0.0386	0.1250 \pm 0.0414
splice	0.0561\pm0.0128	0.0605 \pm 0.0131	tic-tac-t	0.0172 \pm 0.0115	0.0177 \pm 0.0116
vehicle	0.0435 \pm 0.0215	0.0447 \pm 0.0231	vote	0.0471 \pm 0.0333	0.0471 \pm 0.0326
vowel	0.1114 \pm 0.0276	0.1026\pm0.0278	wavef	0.1145\pm0.0136	0.1181 \pm 0.0141
yeast	0.2677\pm0.0344	0.2841 \pm 0.0332	14/27/10		

Table 2: Description of datasets: the number of instances, the number of class, the number of continuous and discrete features

dataset	#inst	#class	#CF	#DF	dataset	#inst	# class	#CF	#DF
abalone	4177	29	7	1	anneal	898	6	6	32
artificial	5109	10	7	–	auto-m	398	5	2	4
auto	205	6	15	10	balance	540	18	21	2
breast-w	699	2	9	–	car	1728	4	–	6
cmc	1473	3	2	7	colic	368	2	10	12
credit-a	690	2	6	9	cylinder	540	2	18	21
diabetes	768	2	8	–	german	1000	2	7	13
glass	214	6	9	–	heart-c	303	2	6	7
heart-h	294	2	6	7	hepatitis	155	2	6	13
house-v	435	2	–	16	hypo	3772	4	7	22
ion	351	2	34	–	iris	150	3	4	–
isolet	7797	26	617	–	kr-vs-kp	3169	2	–	36
letter	20000	26	16	–	lymph	148	4	–	18
magic04	19020	2	10	–	mfeat-f	2000	10	216	–
mfeat-m	2000	10	6	–	mush	8124	2	–	22
musk	476	2	166	–	nursery	12960	2	9	–
optdigits	5620	10	64	–	page-b	5473	5	10	–
pendigits	10992	2	16	–	satimage	6453	7	36	–
segment	2310	7	19	–	shuttle	58000	7	9	–
sick	3372	2	7	22	solar-f	1066	6	–	12
sonar	208	2	60	–	soybean	683	19	–	35
spamb	4601	2	57	–	spect	531	48	100	2
splice	3190	3	–	60	tic-tac-t	958	2	–	9
vehicle	846	4	18	–	vote	435	2	–	16
vowel	990	11	–	11	wavef	5000	3	40	–
yeast	1484	10	8	–					